Introduction

In this report, we analysed various attributes of basketball teams that were part of March Madness brackets to determine how well they would do in the tournament. We intended to discover the best attributes for predicting whether or not a team would pass the first round of the tournament. Initially we wanted to predict if a team would win the whole tournament, but decided that the analysis would become too weighted and unreasonable.

GitHub repo: https://github.com/levinielson/CS6830Project5

PowerPoint: Project 5 Presentation

Dataset

The dataset we obtained came from Kaggle. It needed a little bit of cleaning, as it contained every team that played for given years, and much of the data was a mix between categorical and quantitative data. We classified a 3-point average of over 39 as good, and less than that as bad (in terms of predicting wins and losses). We will assess more of the data cleaning later, but suffice to say that we altered the dataset a bit to fit our needs. We also removed null values from the dataset. We also converted a few values in the dataset from string to float.

Analysis Technique

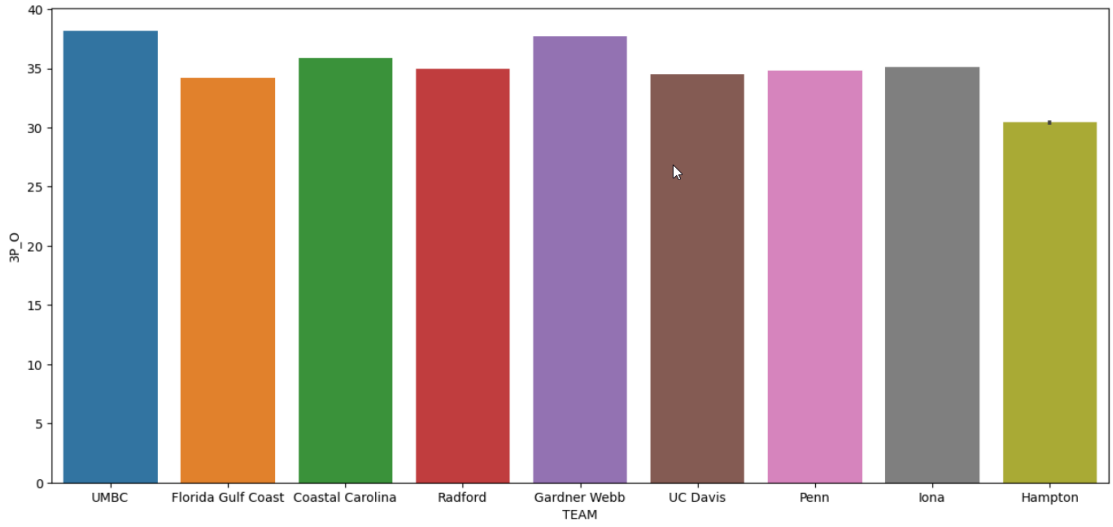
Our main analysis technique was a Naive Bayes Classifier which we trained on a subset of data, then tested and verified on a different portion of the data. Afterwards we also looked at some graphs that talked about the best teams and how they performed in their given year, so we could get a better look at some of the predictors for success in the NCAA tournament. Since the Gaussian Naive Bayes Classifier is a well-known, straightforward, and effective algorithm that is frequently used for classification tasks across a variety of domains, it was selected as the primary analytic method for our dataset.

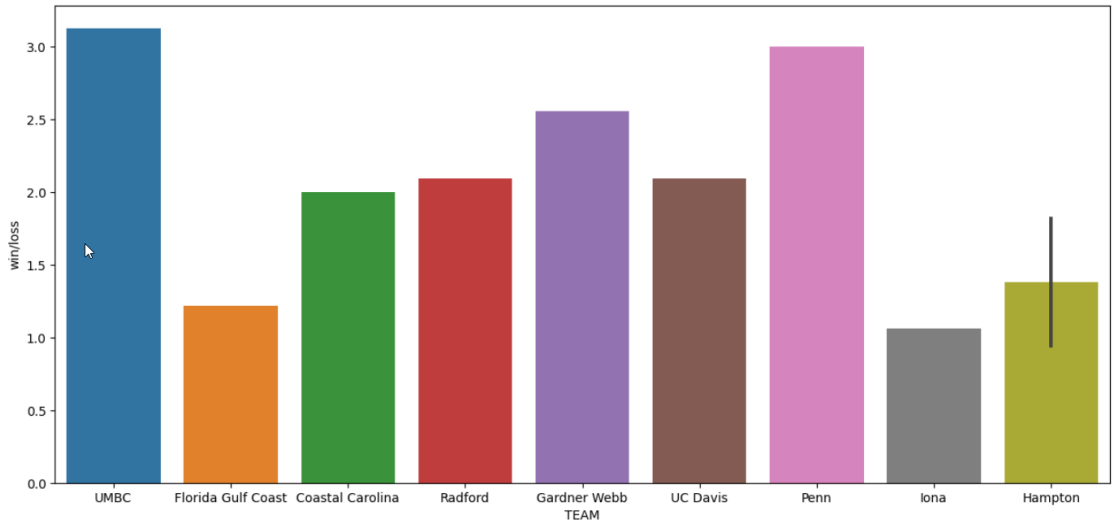
Results

Naive Bayes Classifier:
For our Naive Bayes Classifier, we fed it team data that had several categories. It had a teams seed, whether or not their 3-point percentage was over 39, whether or not their 2-point percentage was over 55, and whether or not their win/loss ratio was greater than 1.5. We experimented with smaller values, and found that the greater the percentage of shots made, the more accurate the classifier was in predicting whether or not a team would pass the first round of the tournament

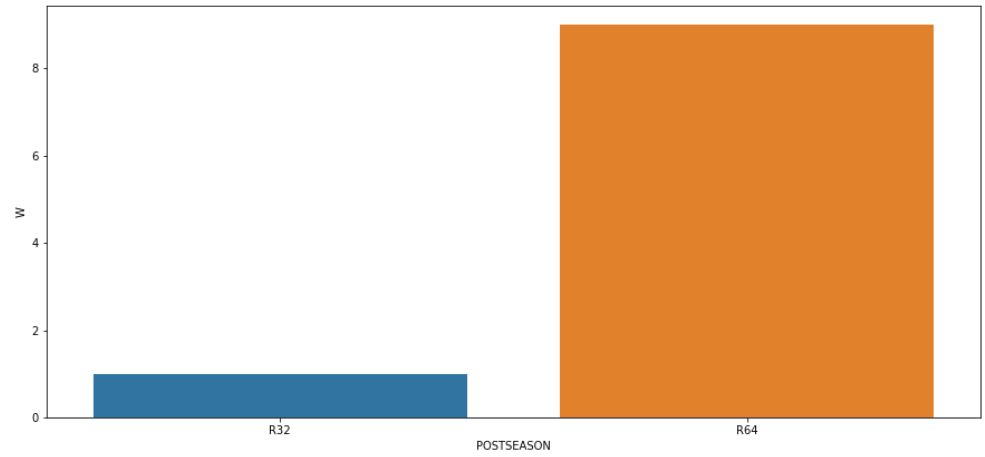
```
best precision_recall_fscore_support value with test size of 0.1 random state of 43 is (
array([0.9047619 , 0.77777778]), array([0.76      , 0.91304348]), array([0.82608696, 0.8
4      ]), array([25, 23]))
Most Informative Features
                seed: 16 = True           First  : Past f =     21.6 : 1.0
                 seed: 2 = True           Past f : First  =     10.2 : 1.0
                seed: 15 = True           First  : Past f =      9.4 : 1.0
                 seed: 3 = True           Past f : First  =      9.3 : 1.0
                 seed: 4 = True           Past f : First  =      8.8 : 1.0
          Win/Loss ratio = 1.75           First  : Past f =      8.2 : 1.0
          Win/Loss ratio = 2.6            Past f : First  =      6.2 : 1.0
            3-point rate = 36.5           Past f : First  =      5.5 : 1.0
            2-point rate = 52.9           Past f : First  =      5.5 : 1.0
          Win/Loss ratio = 3.125          First  : Past f =      4.6 : 1.0
            2-point rate = 49.2           First  : Past f =      4.5 : 1.0
            3-point rate = 35.3           Past f : First  =      4.0 : 1.0
            2-point rate = 50.5           Past f : First  =      4.0 : 1.0
            2-point rate = 53.6           Past f : First  =      4.0 : 1.0
            3-point rate = 32.6           First  : Past f =      3.9 : 1.0
```

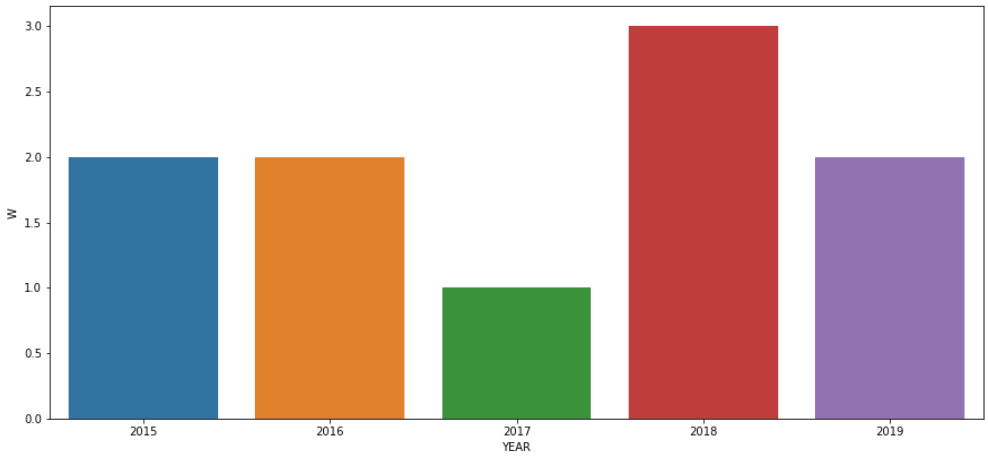
We found that the most informative feature was whether a team was 16th seed. If a team was 16th seed, we discovered that the classifier predicted them to lose the first round 22 times for every one time it predicted them to win it. We were excited about this because it is very accurate, as 16th seed teams rarely pass the first round because they go up against 1st seed teams, the best teams in the nation.





We then took a look at some team's 3-point shooting data and visually compared it with their win/loss ratio. We found that even small variations in the data had large impacts on the teams win/loss ratio, even if it wasn't always a direct impact. This made us more confident in the results that we saw from the classifier.





We see that the teams with seed value of 16 won more matches in 2018 but lost many matches in the start of the tournament.

Technical

We dropped all of the rows that contained NaN data, since the only ones that contained NaN were ones where the team did not get into the march madness bracket. We also cleaned up the seed, 3-point, and 2-point percentage data. We ran the classifier against data that we split, checked the precision, recall, and F-scores, then made graphs that showed some of the teams' ability. We also ran a Gaussian Naive Bayes classifier against the same data as

seen above, and it tended to get better precision for losing teams and recall for winning teams, but didn’t perform as well for F-score